Homework 2: Discovery of Frequent Itemsets and Association Rules

Submit Assignment

Due Monday by 11:59pm **Points** 5 **Submitting** a file upload

The homework can be done in a group of 2 students.

Submission on time, i.e. before the deadline, is awarded with 3 exam bonus points if your homework is accepted, i.e. you have successfully presented and demonstrated your homework to a course teaching assistant during a presentation session.

Introduction

The problem of discovering association rules between itemsets in a sales transaction database (a set of baskets) includes the following two sub-problems [R. Agrawal and R. Srikant, VLDB '94 (http://www.vldb.org/conf/1994/P487.PDF)]:

1. Finding frequent itemsets with support at least ***s***;
2. Generating association rules with confidence at least ***c*** from the itemsets found in the first step.

Remind that an association rule is an implication $X \rightarrow Y$, where X and Y are itemsets such that $X \cap Y = \emptyset$. ***Support*** of the rule $X \rightarrow Y$ is the number of transactions that contain $X \cup Y$. ***Confidence*** of the rule $X \rightarrow Y$ the fraction of transactions containing $X \cup Y$ in all transactions that contain X.

Task

You are to solve the first sub-problem: to implement the Apriori algorithm for finding frequent itemsets with support at least ***s*** in a dataset of sales transactions. Remind that ***support*** of an itemset is the number of transactions containing the itemset. To test and evaluate your implementation, write a program that uses your Apriori algorithm implementation to discover frequent itemsets with support at least ***s*** in a given dataset of sales transactions.

The implementation can be done using any big data processing framework, such as Apache Spark, Apache Flink, or no framework, e.g., in Java, Python, etc.

Optional task for extra bonus

Solve the second sub-problem, i.e., develop and implement an algorithm for generating association rules between frequent itemsets discovered by using the Apriori algorithm in a dataset of sales transactions. The rules must have support at least ***s*** and confidence at least ***c***, where ***s*** and ***c*** are given as input parameters.

Datasets

- As a sale transaction dataset, you can use this **dataset**, which includes generated transactions (baskets) of hashed items – you use any browser, e.g., Google Chrome, or a text editor, e.g., WordPad to view the file under Windows.
- You can also use any other transaction datasets as an input dataset that you can find on the Web.

Readings

- L**ecture 3: Frequent Itemsets**
- **Chapter 6** **(http://infolab.stanford.edu/~ullman/mmds/ch3n.pdf)** of Mining of Massive Datasets, by Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman, 3rd edition, Cambridge University Press, 2020 (**http://www.mmds.org/** **(http://www.mmds.org/)**)
- R. Agrawal and R. Srikant. **Fast Algorithms for Mining Association Rules** **(http://www.vldb.org/conf/1994/P487.PDF)** , VLDB '94

Submission, Presentation and Demonstration

To submit your homework, you upload your solution in a zip file to Canvas. Canvas records the submission time. **Submission on time, i.e. before the deadline, is awarded with 3 exam bonus points if your homework is accepted.** Bonus will not be given, if you miss the deadline. **Your homework solution must include**

1. Source code (with comments);
2. Makefile or scripts to build and run (if needed);
3. Report (in PDF) with a short description of your solution, instructions how to build and to run, command-line parameters, if any (including default values), results, e.g., plots or screenshots.

Within a week after the homework deadline, **you present and demonstrate** your homework on your own laptop to a course instructor. A link to time-slot pool will be provided under Syllabus to **book a time slot for presentation**.

Grading and Bonus Policy

The grade for a homework is **pass/fail**. If you submit report your homework on time and your solution is accepted, you will get **3 bonus points** on your first ID2222 exam whenever you take it. Some homework include **optional tasks for extra bonus**.